THE EVOLUTION OF NON-CODING CHLOROPLAST DNA AND ITS APPLICATION IN PLANT SYSTEMATICS[1]

Scot A. Kelchner[2]

ABSTRACT

This article reviews several proposed mechanisms of molecular evolution operating in non-coding regions of the chloroplast genome and argues that awareness and identification of these mechanisms are essential for improving alignment and phylogenetic analysis of non-coding sequence data. The mechanisms are of five categories: (1) slipped-strand mispairing; (2) insertions and deletions linked with secondary structure formations; (3) inversions associated with hairpins and stem-loop structures; (4) localized or extra-regional intramolecular recombination; and (5) nucleotide substitutions. These mutations seem to be largely a function of sequence structure and pattern and may be highly homoplasious in a parsimony topology; therefore, mutations in non-coding regions of the chloroplast genome are described here as structured, nonrandom, and non-independent events. Established methodologies are based in large part on a collective understanding of genic DNA evolution and may need modification when applied to non-coding sequence data. Here I suggest an approach to the phylogenetic study of non-coding cpDNA that incorporates identification of mutational mechanisms in alignment and homology assessment of indels. I also discuss repercussions of non-coding sequence evolution for such aspects of phylogeny estimation as maximum likelihood, distance, and parsimony analysis, the inclusion of indels as phylogenetic characters, and bootstrapping, jackknifing, and "decay" analysis as measures of clade support.

Key words: alignment, intergenic spacers, introns, molecular evolution, mutational biases, mutational mechanisms, phylogenetic analysis, secondary structure.

There is growing interest in comparative analysis of non-coding chloroplast (non-coding cpDNA) sequences for plant systematic studies at low taxonomic levels. Recognition of the limitations of coding (genic) DNA for resolving relationships at these levels inspired the probing of chloroplast introns and intergenic spacers for phylogenetic utility. Underlying this effort was the reasonable premise that non-coding regions experience limited or no selective pressure and are likely to evolve at rates far surpassing those of genic regions (e.g., Curtis & Clegg, 1984; Wolfe et al., 1987; Palmer, 1987, 1991; Olmstead & Palmer, 1994; Böhle et al., 1994). There was also an expectation that non-coding regions should experience random and independent mutations, both in mode and distribution.

For these reasons, a remarkable number of plant systematics studies currently in progress include a molecular component of comparative analysis of non-coding cpDNA sequences. A considerable amount of work already published has demonstrated the potential phylogenetic utility of discrete non-coding regions in the chloroplast: the *trn*L-*trn*F spacer (e.g., Gielly & Taberlet, 1994; Mes & t'Hart, 1994; van Ham et al., 1994; Sang et al., 1997; Cros et al., 1998; Bayer & Starr, 1998), the *trn*T-*trn*L spacer (Böhle et al., 1994, 1997; Small et al., 1998), the *rpo*A-*pet*D and *rps*11-*rpo*A spacers (Peterson & Seberg, 1997), the *atp*B-*rbc*L spacer (Golenberg et al., 1993; Hodges & Arnold, 1994; Natali et al., 1995; Samuel et al., 1997; Savolainen et al., 1997; Setoguchi et al., 1997; Hoot & Douglas, 1998), the *rbc*L-*psa*I spacer (Morton & Clegg, 1993), the *psb*A-*trn*H spacer (Aldrich et al., 1988; Sang et al., 1997), the *acc*D-*psa*I spacer (Small et al., 1998), the *rpl*16-*rpl*14 and *rps*8-*rpl*14 spacers (Wolfson et al., 1991), the intron surrounding *mat*K (Johnson & Soltis, 1994), the *rpo*C1 intron (Downie et al., 1996a, 1996b; Asmussen & Liston, 1998; Downie et al., 1998), the *rpl*16 intron (Jordan et al., 1996; Kelchner, 1996; Kelchner & Clark, 1997; Schnabel & Wendel, 1998; Baum et al., 1998; Small et al., 1998), the *trn*L intron (Sang et al., 1997; Bayer & Starr, 1998; Kajita et al., 1998; Bay-

[1] I thank Conny Asmussen, Lyn Cook, Steve Dickie, Jonathan Wendel, Rich Cronn, and Randy Small for many interesting and insightful discussions on the topic of non-coding sequence evolution; R. G. Olmstead for his valuable comments on an earlier version of this paper; R. J. Bayer, M. D. Crisp, J. G. West, C. L. Kelchner, and B. Oxelman for review of the manuscript; Victoria Hollowell for her careful editing and commentary; and the anonymous reviewers, whose insightful comments contributed to the clarity and presentation of ideas in this paper.

[2] Centre for Plant Biodiversity Research, Australian National Herbarium, CSIRO Plant Industry, G.P.O. Box 1600, Canberra, A.C.T., 2601, Australia. Second affiliation: Division of Botany and Zoology, The Australian National University, Canberra, Australia, scot.kelchner@pi.csiro.au.

ANN. MISSOURI BOT. GARD. 87: 482–498. 2000.

er et al., 2000), the *rps*16 intron (Liden et al., 1997; Oxelman et al., 1997), and the *ndh*A intron (Small et al., 1998).

The literature above not only reveals profound differences between the evolution of non-genic and genic cpDNA, but critically contradicts initial assumptions of constraint-free evolution in non-coding regions. Recurring difficulties associated with non-coding sequence data include alternative alignment possibilities of insertions and deletions (indels), regions of length mutation in which homology assessment is questionable or impossible, and the occurrence of localized "hot spots" of inferred excessive mutation, frequently to the point of saturation and loss of phylogenetic signal. How best to proceed with the phylogenetic analysis of such regions should be a topic of considerable concern (see Golenberg et al., 1993; Downie et al., 1996a; Kelchner & Clark, 1997; Sang et al., 1997; Downie et al., 1998).

It is now evident that sequence evolution in non-coding regions of the chloroplast is far more complex than previously supposed. Both introns and intergenic spacers are thought to embody a considerable degree of sequence structure, sometimes in a manner similar to that of ribosomal DNA (rDNA). This structure may generate either regionalized sequence conservation or mutational hot spots of both nucleotide substitutions and insertion/deletion events. Sequence-directed initiators of mutational events may persist as "mutational triggers" (Kelchner, 1996; Kelchner & Clark, 1997), dramatically increasing the possibility of reversal or parallel gain of mutations, particularly length mutations or minute inversions. Hence, there exist essential violations of the assumptions of randomized and independent character evolution embedded in much of the current phylogenetic methodology for comparative sequence analysis—methodology that is based largely on observational comparative study of coding sequence data. Considering that these are today's commonly employed tools for phylogeny estimation based on DNA sequences, there has been as yet remarkably little controversy in the literature about their application to non-genic sequence data.

There are ways to account for mutational patterns observed in non-coding DNA. Comparative studies of non-coding cpDNA sequences during the past decade in particular (e.g., Palmer, 1985; Blasko et al., 1988; vom Stein & Hatchel, 1988; Wolfson et al., 1991; Golenberg et al., 1993; Gielly & Taberlet, 1994; Morton, 1995a; Downie et al., 1996a; Kelchner & Wendel, 1996; Kelchner & Clark, 1997; Sang et al., 1997) have allowed inference of specific underlying mutational mechanisms responsible for generating sequence diversity in non-coding regions of the chloroplast genome. Unfortunately, these mechanisms are often invoked, but rarely incorporated, into the analysis.

Recognition of the potential of structured molecular evolution in non-coding cpDNA regions to improve alignment and assessment of phylogenetic relationships is, I believe, critical for the development of functional molecular systematic research based on non-coding sequence data. Toward this end, I endeavor here to illustrate the following: (1) non-coding regions are highly structured and their elements evolve non-randomly and non-independently; (2) this structure may be used to align the sequence matrix and better assess homology; (3) the resulting gaps in the aligned matrix may contain phylogenetically important information and should be used in a phylogenetic analysis; and (4) the mode of non-coding sequence evolution described here may have potentially serious repercussions for the accuracy of genetic-distance, maximum likelihood, and parsimony analyses, and for bootstrapping and jackknifing techniques. A description of proposed mechanisms of non-coding sequence evolution is followed by a discussion of the appropriateness of current alignment and analysis procedures, with the expectation that it may provide a more informed approach to the application of non-coding sequence data in plant systematics research.

This article is not intended to be a complete review of literature pertaining to the evolution of introns and intergenic spacers in all genomes of an organism. Instead, it serves as a brief review of current literature on non-coding cpDNA regions, and summarizes mutational mechanisms suggested to occur in these regions. Discussed are some of the serious implications this manner of molecular evolution has for the assumptions underlying models employed today by plant molecular systematists.

Mechanisms of Non-Coding Sequence Evolution

The strength of any phylogenetic estimation rests on the accuracy of character homology assessment. Thus, the molecular systematist strives to maximize character homology by the careful alignment of DNA sequences in a data matrix. Fundamental to any alignment procedure of non-coding cpDNA sequence data should be a familiarity with mutational mechanisms directing molecular evolution in non-coding regions. Recognition of these mechanisms as generators of specific mutations can be a powerful tool for the placement of gaps and for the assessment of probable homology of insertions and deletions (Kelchner, 1996; Kelchner & Clark, 1997).

SLIPPED-STRAND MISPAIRING (SSM)

A widely reported mechanism of length mutation in non-coding regions of the chloroplast is slipped-strand mispairing (SSM). SSM is thought to be a major, even principal, factor in length mutations within non-coding regions of the chloroplast, mitochondrial, and nuclear genomes (e.g., Levinson & Gutman, 1987; Hancock, 1995; Wolfson et al., 1991; Kelchner & Clark, 1997; Sang et al., 1997). Length mutations are important components of non-coding sequence evolution and have been suggested to occur at least as frequently as base substitutions in some chloroplast non-coding regions (Curtis & Clegg, 1984; Wolfe et al., 1987; Zurawski & Clegg, 1987; Clegg & Zurawski, 1992; Golenberg et al., 1993; Gielly & Taberlet, 1994; Clegg et al., 1994).

Slipped-strand mispairing is thought to proceed by a localized mispairing of single-stranded DNA in regions of sequence repeats, as either a string of mononucleotide repeats or tandemly arranged multibase repeat units (Palmer, 1991; Wolfson et al., 1991; Cummings et al., 1994; Hancock, 1995; reviewed by Levinson & Gutman, 1987). Diagrams of proposed SSM mechanics can be found in Levinson and Gutman (1987) and Wolfson et al. (1991). Because A/T-rich regions of bacterial genomes are particularly susceptible to slipped-strand mispairing (Levinson & Gutman, 1987), one could expect a similar effect in the A/T-rich non-coding regions of the chloroplast genome (Wolfson et al., 1991). This is not to imply that SSM acts uniquely on A and T nucleotides; aligned non-coding sequence matrices often infer inserted repeats containing G and C nucleotides, sometimes as pure strings of G or C mononucleotide repeats.

Strings of mononucleotide repeats, particularly of A or T, appear frequently in non-coding cpDNA, and slipped-strand mispairing may potentially generate length mutations within these strings. The difficulty in assessing homology of length variation in long strings of repeats, whether mononucleotide or multinucleotide repeats, derives from the increasing potential for further length mutation relative to string length (Streisinger & Owen, 1985; Golenberg et al., 1993; Kelchner & Clark, 1997; Sang et al., 1997). Subsequent SSM activity may either generate additional repeats of the initial sequence or delete sequence susceptible to slipped-strand mispairing. Perhaps an equilibrium might exist between the probability of inserting subsequent length mutations and the probability of removing sequence from the repeat string. Whether such an equilibrium is present or not, there may be a competitive phenomenon that keeps the length of tandem repeated sequence units continually in flux. Representation of long repeat strings in non-coding sequence alignments would therefore be a "snapshot" of sequences experiencing continual insertions and deletions at that locality.

It follows that a point substitution within a long string of mononucleotide repeat units could act as a stabilizing factor, disrupting its previous uniformity and lowering the probability of further SSM events. Such a substitution would directly influence ensuing mutations in the region and is one example of a non-independent character mutation in non-coding DNA. If the situation were reversed, with a non-homogeneous sequence becoming a string of repeat units, the likelihood of an SSM event would increase and could induce further non-independent mutations by the addition or removal of repeated sequence by slipped-strand mispairing.

As an aid to alignment, SSM-generated insertions and deletions can be used to position and determine number of gaps. A quick study of a repeat unit or the flanking sequence of a gap may be enough to determine if slipped-strand mispairing is the likely progenitor of an observed length mutation. Occasionally, evidence of an SSM event may not be apparent, particularly if a deleted sequence is not a direct repeat of its flanking sequence, or if a subsequent length mutation due to another mechanism obscures an earlier SSM event (Kelchner, 1996).

STEM-LOOP SECONDARY STRUCTURE

Striking to both intergenic spacers and introns in the chloroplast genome is the presence and number of probable secondary structures referred to as "stem-loops." Stem-loops are believed to occur during single-stranding events when inverted repeats meet to form a region of pairing (the stem) surmounted by their interceding sequence (the loop). Such structures have been widely discussed for ribosomal DNA, with ITS and 18S rDNA regions being of particular interest to the plant systematist (see Baldwin et al. (1995), Soltis et al. (1997), and Soltis & Soltis (1998) for discussion of secondary structures in these regions and their phylogenetic implications).

Probable stem-loop secondary structure is commonly reported in non-coding regions of organellar genomes (e.g., Michel et al., 1989; Buroker et al., 1990; Golenberg et al., 1993; van Ham et al., 1994; Gielly & Taberlet, 1994; Natali et al., 1995; Rigaa et al., 1995; Downie et al., 1996b; Kelchner & Wendel, 1996; Kelchner & Clark, 1997; Sang et al., 1997; Downie et al., 1998). Gielly and Taberlet (1994) reported several probable stem-loops in the *trn*L-*trn*F region of the chloroplast genome, including nine highly probable structures within the *trn*L intron itself. All other introns in the chloroplast genomes of land plants are classified as Group II introns and share a diagnostic secondary structure of six well-defined stem-loop domains (Kohchi et al., 1988; Michel et al., 1989; Downie et al., 1996b; Downie et al., 1998). Diagrams of putative single-stranded secondary structure of introns may be found in Michel and Dujon (1983), Michel et al. (1989), and Downie et al. (1998).

Loop regions of stem-loop secondary structures are often associated with hot spots for mutation in non-coding regions, both of nucleotide substitutions and indel events (vom Stein & Hatchel, 1988; Aldrich et al., 1988; Golenberg et al., 1993; Gielly & Taberlet, 1994; van Ham et al., 1994; Clegg et al., 1994; Ferris et al., 1995; Downie et al., 1996b; Kelchner & Clark, 1997). Indels located in probable loop sequence are frequently inserted or deleted repeat units likely the result of SSM. However, length mutations not attributable to slipped-strand mispairing often occur within loop sequences as well and may be remnants of recombination events.

Although indels are most common in the terminal loop, they may occur anywhere along a secondary structure. For example, Kelchner and Clark (1997) detected what appeared to be an entire deletion of a small sub-loop positioned partway up the stem of an *rpl*16 intron stem-loop in *Oryza sativa*. Such side loops, when present, may be removed in some taxa without compromising the favorability of a stem formation. Occasionally, small segments of the stem itself will be deleted, decreasing the stem length, though perhaps not to an extent that would annihilate possible secondary structure formation.

Very large loops are often associated with regions of chaotic or "labile" length variation characteristic of many non-coding cpDNA sequence matrices (e.g., Golenberg et al., 1993; Downie et al., 1996a; Soltis et al., 1996; Kelchner & Clark, 1997; Baum et al., 1998). Homology assessment here can be difficult or impossible, and the conservative approach of removing these regions from the data matrix before phylogenetic analysis is frequently adopted.

In contrast to the loop of stem-loop secondary structures being highly susceptible to nucleotide substitutions and length mutation, the inverted repeated sequence composing the stem is frequently conserved in character (Learn et al., 1992; Gielly & Taberlet, 1994; Downie et al., 1996a, 1996b; Kelchner & Clark, 1997), particularly when stems are long and possess highly favorable energy of formation values (ΔG values; see Kelchner & Wendel, 1996; Dumolin-Lapègue et al., 1998). A sequence involved in stem formation is less available for substitution and length mutation because it is paired with its sister repeat; this can engender non-randomly and non-independently evolving sequence units.

Similar to ribosomal RNA and rDNA secondary structure (e.g., Curtiss & Vournakis, 1984; Wheeler & Honeycutt, 1988; Dixon & Hillis, 1993; Soltis & Soltis, 1998), a nucleotide substitution occurring in a stem sequence of a non-coding cpDNA region could compromise secondary structure formation. Compensatory mutation may then occur to preserve the potential for structure formation (Kelchner, 1996; Kelchner & Clark, 1997). Although sequence conservation may be present merely as a function of sequence pattern (pèrhaps the case in intergenic spacers), the degree of secondary structure conservation in a chloroplast Group II intron suggests secondary structures are integral to proper functioning of the intron (Clegg et al., 1986; Learn et al., 1992; Downie et al., 1996a). Experimental evidence has shown some of this structure is essential for auto-splicing mechanisms in Group I and II introns (Bonnard et al., 1984; Kohchi et al., 1988; Dujon, 1989; Cech, 1990; Michel & Westhof, 1990; Hibbett, 1996).

Identification of probable secondary structure can be valuable when aligning and analyzing non-coding sequences by improving gap positioning and the appraisal of character homology. Gaps flanked by inverted repeats and regions relatively rich in G and C content are suspect as possible stems of secondary structures. As noted, regions of chaotic length mutations are correlated with loops, so the boundaries of a chaotic region will frequently correspond with inverted repeats that can form a stem, even if they do not directly neighbor the chaotic region. Computer programs such as OLIGO (Rychlik & Rhoads, 1989), MULFOLD (Jaeger et al., 1989; Zuker, 1989), and GCG's Stemloop (Genetics Computer Group, Madison, Wisconsin) can assist in the detection of secondary structure in non-coding sequences. A search can be conducted by hand, particularly if a published data set exists for the region. Free energy of formation values (ΔG) can be calculated with some of the prior software as an appraisal of the likelihood of formation of a particular secondary structure (see Kelchner & Wendel (1996) for an example where ΔG values were applied to parallel inversion events in their data).

MINUTE INVERSIONS

Minute inversions of four to six base pairs have been linked to small stem-loop secondary structures commonly referred to as hairpins (Kelchner & Wendel, 1996). Hairpins consist of a stem composed of nearly adjacent inverted repeats producing a stem-loop structure with a particularly small loop. This loop may become inverted by recombination, and the inversion may be so small that it either escapes notice during alignment (Kelchner & Wendel, 1996; Kelchner & Clark, 1997), or the inverted sequence matches particular bases of the uninverted sequence, resulting in a confusing array of minute gaps (see Golenberg et al., 1993).

Identifying minute inversions can require careful attention when aligning sequence data, particularly if alternative gap weighting schemes of an alignment program have not been rigorously explored. Candidates for a hidden inversion are several adjacent nucleotide substitutions, a series of tiny gaps, or a gap that demonstrates no repeat aspect to its sequence structure. Alternatively, one could investigate these probable secondary structures by hand or with a secondary structure computer program. Failure to recognize minute inversions in a sequence data set has several repercussions for phylogenetic analysis, discussed fully in Kelchner and Wendel (1996) and summarized here in Analysis of Non-Coding Sequence Data.

Finally, small inversions associated with hairpins may be highly susceptible to reversal and parallelism within a study group, even at the interspecific level (Kelchner & Wendel, 1996; Kelchner & Clark, 1997; Sang et al., 1997; Dumolin-Lapègue et al., 1998). This susceptibility to reversal or parallelism is due to the persistence of the mutational trigger (Kelchner & Clark, 1997)—the nearly adjacent inverted repeats—after the initial inversion event.

NUCLEOTIDE SUBSTITUTIONS

Nucleotide substitutions are generally reported as being more common in non-coding than in coding regions (Wolfe et al., 1987; Zurawski & Clegg, 1987; Olmstead & Palmer, 1994; Hoot & Douglas, 1998; however, see Sang et al., 1997, for an exception). Surprisingly, a number of studies report nucleotide substitutions as being just equal to or less frequent than length mutations in closely related taxonomic groups (Curtis & Clegg, 1984; Wolfe at al., 1987; Zurawski & Clegg, 1987; Clegg & Zurawski, 1992; Golenberg et al., 1993; Gielly & Taberlet, 1994; however, see Small et al., 1998).

Percent AT content is quite variable in non-coding cpDNA regions, though it is generally higher than the average value for the chloroplast genome (Shimada & Sugiura, 1991; Downie at al., 1996a; Small et al., 1998). Because of their high AT content, non-genic regions must make a significant contribution to the high overall frequency of A and T in the chloroplast genome. Kajita et al. (1998) reported an AT content of 67% in the *trn*L-*trn*F spacer and *trn*L intron, Kelchner and Clark (1997) reported 70.5% AT composition in the intron of chloroplast gene *rpl*16 in bamboos, and Small et al. (1998) found an incredible 77.1% AT content in the intergenic spacer *trn*T-*trn*L in *Gossypium*. Undoubtedly, this unequal tendency toward AT richness in non-genic chloroplast DNA has several as yet undetermined implications for phylogenetic analysis of non-coding sequence data. At a minimum, it introduces a strong base composition bias into the analysis.

Substitutions may demonstrate rather high levels of homoplasy in non-coding cpDNA regions due to the frequency of inferred multiple-hit sites (nucleotide sites experiencing multiple substitution events). Multiple-hit sites occur even at very low estimates of percent sequence divergence (Kelchner, 1996; Kelchner & Clark, 1997), suggesting that the accepted coding region estimates of "around 10–15%" sequence divergence for optimal phylogenetic signal may be inadequate measures for phylogenetic utility of a non-coding region.

Precise understanding of mechanisms underlying multiple-hit substitutions in non-coding DNA is lacking. However, attributes of the molecular evolution of non-coding regions influence the manner of nucleotide mutation or the distribution of nucleotide substitution events in an intron or intergenic spacer. Stem sequence and loop regions may differentially permit mutations, resulting in non-randomly distributed and non-independent nucleotide substitutions. Statistical significance of differential mutation rates in loops relative to stems may be tested for an adequate distribution model (see Olmstead et al.'s (1998) test for stochastic mutation in the chloroplast genes *ndh*F and *rbc*L), yet has rarely, if ever, been performed on non-coding cpDNA data sets.

In addition to secondary structure affecting the random distribution of nucleotide substitutions, there may be constraints on the type of mutation an individual site experiences. For example, there is a correlation between transition/transversion ratios and neighboring base composition in non-coding regions (Morton, 1995a, b; Morton et al., 1997; Savolainen et al., 1997). The correlation suggests that nucleotides flanked by A and/or T will demonstrate a significant tendency toward transversion mutations. Such a tendency limits possible nucleotide replacements at these sites, increasing the chance of parallelism and reversals, particularly if the site experiences multiple hits. One would also expect transversion substitutions to be more common in data sets of high AT content.

INTRAMOLECULAR RECOMBINATION

Intramolecular recombination on an extra-regional or genomic scale has been suggested between adjacent or nearby repeats in the chloroplast genome (Howe, 1985; Palmer et al., 1985; Palmer et al., 1987; Blasko et al., 1988; Ogihara et al., 1988; Milligan et al., 1989; Kanno & Hirai, 1992; Kanno et al., 1993; Morton & Clegg, 1993; Hoot & Palmer, 1994). In the context of non-coding sequence comparison, such a large-scale recombination involving the particular region of study could result in indels of surprising size that contain sequence content not readily identifiable in origin.

Recombination events may operate on a finer scale within a discrete non-coding region. Occasionally one infers extensive deleted sequence in an alignment with no apparent mechanistic explanation, presence of a small or moderately sized inversion, or a large insertion showing little congruence with surrounding sequence pattern. Such mutations suggest intramolecular recombination, and they frequently occur in the loop regions of probable secondary structures. Sequences involved in stem-loops may be particularly susceptible to recombination events due to the conserved inverted repeats and mutationally flexible loop. Therefore, such structures could experience interactive recombination with other stem-loops, particularly with those existing in complementary sequence position.

Recombination involving the entire loop of a secondary structure may occur, particularly in structures with long stems, resulting in minute or moderate-sized inversions in both intron and intergenic spacer regions (Natali et al., 1995; Kelchner & Wendel, 1996; Kelchner & Clark, 1997; Sang et al., 1997). Such incidents are often homoplasious (Kelchner & Wendel, 1996; Kelchner & Clark, 1997; Sang et al., 1997; Dumolin-Lapègue et al., 1998) due to the persistence of the mutational trigger; in this case, the hairpin stem.

Intramolecular recombination is a notable alternative to slipped-strand mispairing as a source for certain inserted or deleted tandem-repeat length mutations (Palmer, 1985; Blasko et al., 1988). However, Wolfson et al. (1991), Sang et al. (1997), and Kelchner and Clark (1997) suggested SSM is a more likely mechanism for length mutation in their studies of chloroplast introns and intergenic spacers.

ALIGNMENT

There are many philosophies for sequence alignment, and much of the literature centers on the proper application of computer software for this purpose. The structure present in a non-coding cpDNA sequence makes it an excellent example for discussing what I believe to be the fundamental problem of most computer alignment programs: defining the nucleotide as a discrete and independent character. The identification of secondary structure and mutational mechanisms in the data may greatly improve on current algorithmic alignments of gaps, and thus on assessment of character homology.

Many have found software, particularly versions of CLUSTAL (Higgins et al., 1992; Thompson et al., 1994), to be of help at least initially with the alignment of non-coding sequences. The alignment is then subjected to an "improvement by hand" to position gaps (e.g., Samuel et al., 1997; Downie et al., 1998; Bayer & Starr, 1998; Kajita et al., 1998). This procedure saves time if the sequences are similar in length, but when indels become numerous in the data matrix the difficulties of alignment dramatically increase. This is because most alignment software initially regards each character in the matrix as an independent unit, unless otherwise specified by particular position or gap weighting schemes defined by the user. The software is incapable of determining when mutations other than substitutions have arisen, such as non-independent insertions, deletions, or inversions correlated with SSM and secondary structure. Appropriate weighting for these mutations that could be incorporated into an alignment algorithm is, at present, undeveloped.

The Elision method of Wheeler et al. (1995) attempts to improve gap placement and indel homology by alignment software. The Elision method uses standard alignment algorithms to produce a series of competing alignments based on varying gap weighting schemes. These competing alignments are then combined in a single matrix and an analysis is performed, with the effect that support is increased for aligned regions that most frequently appear among the various gap-weighting schemes. This method aims at objectivity, but makes no improvement on the alignment algorithm's inability to assess mutation types other than independent point substitutions. Mutations in non-coding regions are influenced by surrounding sequence structure and frequently occur not as independent base mutations but as linked multinucleotide mutation events, like the insertion of a repeat unit (Kelchner, 1996; Kelchner & Clark, 1997). The likelihood that many non-coding mutations are derived from sequence fragments that are inserted, deleted, inverted, or otherwise rearranged, negates the assumption of discrete, independent nucleotide characters underlying all alignment algorithms, as well as any extension of those algorithms like the Elision method.

At a minimum, those using sequence alignment programs to establish putative homology of characters in their data matrix should experiment with a wide variety of gap-weighting options. These options, however, may not reveal the underlying mutational mechanisms occasioning sequence rearrangements in chloroplast non-coding regions. They may, however, facilitate the rapid alignment of segments of the matrix that share consistent sequence integrity and thus pinpoint regions of variable length that require special consideration.

Alternatively, some have avoided alignment programs entirely and describe aligning sequences by hand (e.g., Golenberg et al., 1993; Hodges & Arnold, 1994; Kelchner & Clark, 1997). This approach facilitates a careful study of the matrix as it forms and increases the researcher's familiarity with mutations in the sequences. However, alignment by hand, especially when dealing with considerably divergent taxa or with the presence of a great number of length mutations, can be tedious and time consuming.

Kelchner and Clark (1997) suggested that awareness of the proposed mutational mechanisms active in non-coding regions can be useful for inferring and positioning gaps and ultimately in assessing homology. Golenberg et al. (1993) were the first to detail a criterion for aligning gaps in non-coding cpDNA matrices. Based on their example, Kelchner (1996) and Kelchner and Clark (1997) modified the alignment criterion for chloroplast *rpl*16 intron sequences. Hoot and Douglas (1998) also revised Golenberg et al.'s (1993) method of gap alignment, framing the beginnings of a nomenclatural procedure for defining gap categories. Although a nomenclatural system is not requisite for gap treatment in a phylogenetic analysis, it may be useful in collating information of inferred mutational mechanisms if universally applied in non-coding DNA studies.

ALIGNMENT ISSUES: EXAMPLES FROM NON-CODING cpDNA DATA

Here I present examples (Kelchner & Wendel, 1996; Kelchner & Clark, 1997; Kelchner, unpublished data) to illustrate the inference of mutational mechanisms in non-coding cpDNA sequences and demonstrate the practice of applying mechanistic explanations to alignment and homology assessment. Nucleotides in lower-case bold print are inferred insertions; underlined nucleotides indicate the probable progenitor sequence of an insertion or, in Examples 4 and 5, call attention to a particular sequence of interest.

A common type of insertion in non-coding cpDNA is a direct repeat of a neighboring sequence ("Type 1a" gap; Golenberg et al., 1993; Hoot & Douglas, 1998). These often take the form of variable-length strings of a mononucleotide repeat unit (Example 1).

EXAMPLE 1.

```
1. TTAAAAAAAAA---TTGA
2. TTAAAAAAAAAA--TTGA
3. TTAAAAAAAA----TTGA
4. TTAAAAAAAAAAAATTGA
```

Homology can be highly uncertain for these repeated nucleotides. Therefore, such regions are either removed from consideration as potential phylogenetic characters (a conservative approach) or included as coded gap characters corresponding to length of the repeat string (often becoming highly homoplasious in the context of a resulting topology). Uncertainty of homology is exacerbated by potential inaccuracies of enzymatic processes during PCR amplification and sequencing, which can also generate variable-length repeat strings independent of the template's sequence constitution. When strings of adjacent mononucleotide repeats are highly variable in length in a matrix and reach or exceed the range demonstrated above, they become more likely to experience further SSM mutation. For this reason, it is perhaps most reasonable to remove such areas from consideration in a phylogenetic analysis.

Insertions can also be multinucleotide repeat units of a neighboring sequence, as demonstrated in Example 2 by the inserted repeat unit **ataaa** ("Type 1b" gap; Golenberg et al., 1993; Hoot & Douglas, 1998).

EXAMPLE 2.

1. ATAAAACAAA-----GAGCG
2. ATAAAATAAA**ataaa**GAGCG
3. ATAAAATAAA-----GAGCG
4. ATAAAATAAA-----GAGCG

An inserted repeat of this nature could be extensive in length and may be difficult to recognize as a repeat unit during alignment (for example, I have identified a 73 bp inserted repeat [unpublished data] in the *trn*T-*trn*L intergenic spacer in Myoporaceae). A repeat unit by its very nature shares nucleotide content and order with flanking sequence; therefore, multiple gaps may be inferred by pairing segments of an inserted repeat with its progenitor sequence. This is particularly problematic if the insertion or its progenitor has experienced subsequent nucleotide substitutions.

Even when a single gap is inferred, positioning of the gap may hide evidence that the insertion is a repeat unit. Example 3 is reproduced from Kelchner and Clark (1997) and demonstrates how a repeat unit may be obscured in a sequence matrix.

EXAMPLE 3.

A.

1. GGTTATGA ----- ATTAACA
2. GGTTATAA ----- ATTAACA
3. GGTTATAA **tataa** ATTAACA
4. GGTTATAA **tataa** ATTAACA

B.

1. GGTTAT-- ---GA ATTAACA
2. GGTTAT-- ---AA ATTAACA
3. GGTTATAA **tataa** ATTAACA
4. GGTTATAA **tataa** ATTAACA

C.

1. GGTTA--- --TGA ATTAACA
2. GGTTA--- --TAA ATTAACA
3. GGTTATAA **tataa** ATTAACA
4. GGTTATAA **tataa** ATTAACA

Alignment possibilities A, B, and C were equally probable using CLUSTAL W (Thompson et al., 1994). Only alignment A reveals the insertion is a repeat unit—a common mutation type in non-coding regions. If alignment options B or C were used for phylogenetic analysis, the content of the insertion would be of unexplainable origin (though still possible) and the potential of incorrectly assessing nucleotide homology in the region may be considerable.

Any of the gap positions in this particular example would not affect a topology generated from these four taxa, but gap positioning may have a significant effect in a larger matrix of more distantly related taxa. The position of the gap in alignment 3A and detection of the repeat unit may also be relevant in determining a weighting scheme for these non-independent characters.

Length mutations may overlap with one another to create a progressive-step indel. In the more extreme cases, appraisal of homology in these regions can be very difficult, or impossible (Palmer et al., 1985; Downie et al., 1996b; Kelchner & Clark, 1997). Example 4 demonstrates a probable progressive-step indel in which two possible placements exist for the repeat TTGA. Note that the underlined sequence is a direct repeat of the preceding sequence TCGTAATTGA in the matrix.

EXAMPLE 4.

1. AATCGTAATTGA ---------- ----AACAGA
2. AATCGTAATTGA ---------- ----AACAGA
3. AATCGTAATTGA TCGTAA**TTGA** ----AACAGA
4. AATCGTAATTGA TCGTAA**TTGA** ----AACAGA
5. AATCGTAATTGA TCGTAA**TTGA** **ttga**AACAGA

If part of the underlined **TTGA** in sequences 3 and 4 is moved from its current position to align with **ttga** in sequence 5, the possibility that the **ttga** sequence is a direct repeat of the preceding sequence may be obscured; however, this alignment choice would not be impossible. As the preceding sequence to the underlined 10 bp repeat does not contain this additional **ttga** repeat, we can infer that two separate events have given rise to an initial 10 bp insertion in sequences 3 and 4, followed by an additional 4 bp insertion in sequence 5. Whether **ttga** itself or the preceding **TTGA** is the subsequent inserted mutation is impossible to determine. In this case, either alternative alignment of the TTGA unit would cause no effect in a phylogenetic analysis; it is most important here to discern the two length mutation events. If any potentially informative nucleotide substitutions were present in either of the repeat units in Example 4, these substitutions should be excluded from a phylogenetic analysis on the basis that nucleotide homology of the repeats is not discernable.

The example above suggests that homology may be indicated by the length of insertions or deletions in a gap, although such an assumption is not without risk. Example 5 below demonstrates multiple possible alignments of the **gatt** repeat unit (represented individually by sequences 2, 3, and 4) with the insertion in sequence 1.

EXAMPLE 5.

1. CAGATTGATTGATTATTATACTGATTATGC
2. CAGATT----------------**gatt**ATGC
3. CAGATT**gatt**----------------ATGC
4. CAGATT----**gatt**------------ATGC
5. CAGATT--------------------ATGC

Again, actual homology is impossible to assess with confidence, for there exist three GATT repeat units in the insertion in sequence 1. In cases like this, homology is often inferred on the basis of length of indel and minimum number of gaps required to position the repeat. Hence, the **gatt** repeats in sequences 2, 3, and 4 would be aligned one above the other and on one side of the gap to reduce the number of inferred indel events. When coding indels as characters, this would be a reasonable solution in lieu of other evidence for indel origin, and the repeat **gatt** would be treated as homologous for those sequences that contain it.

Equal length of insertions may not be strong evidence of their homology (Kelchner, 1996; Kelchner & Clark, 1997; Hoot & Douglas, 1998). Consider the insertions in Example 6A.

EXAMPLE 6A.

1. GGTTAAT **tctat** TCTATCT
2. GGTTAAT **ttaat** TCTATCT
3. GGTTAAT ttaat TCTATCT
4. GGTTAAT ----- TCTATCT
5. GGTTAAT ----- TCTATCT

Alignment of the insertions in Example 6A results in the probably mistaken homology of indels in sequences 2 and 3 with that of sequence 1. The insertion in sequence 1 likely arose from an inserted repeat of the sequence to the right of the gap, TCTAT. This would be a more parsimonious explanation, in terms of total number of mutation events, than to infer a single inserted repeat followed by two adjacent nucleotide substitutions in sequence 1. Sequences 2 and 3 probably share a similar origin as a repeat of the preceding sequence TTAAT. The events, aligned as they are in Example 6A, are probably non-homologous. A re-alignment could be performed to accommodate the two separate indel events (Example 6B), even though the insertions are of the same length and the alignment infers an additional gap (see Hoot & Douglas, 1998).

EXAMPLE 6B.

1. GGTTAAT ----- tctat TCTATCT
2. GGTTAAT ttaat ----- TCTATCT
3. GGTTAAT ttaat ----- TCTATCT
4. GGTTAAT ----- ----- TCTATCT
5. GGTTAAT ----- ----- TCTATCT

There is a hazard that minute inversions (Kelchner & Wendel, 1996) can be completely obscured in a matrix if they introduce no gaps during alignment, particularly if alternative gap-weighting schemes have not been rigorously pursued. If present and unrecognized in a data matrix, minute inversions may overweigh a particular mutation by interpreting the single mutation event (an inversion) as multiple apomorphies of adjacent nucleotide substitutions. Example 7 below illustrates a situation in which sequences 2 and 3 share the inversion **TTGG** to **CCAA** (from Kelchner & Wendel, 1996).

EXAMPLE 7.

1. TAATATT **TTGG** AATATTA
2. TAATATT **CCAA** AATATTA
3. TAATATT **CCAA** AATATTA
4. TAATATT **TTGG** AATATTA
5. TAATATT **TTGG** AATATTA

If the inversion is of sufficient length to introduce multiple gaps in the matrix (see Golenberg et al., 1993; Sang et al., 1997), two possibilities can occur: the gaps will be misaligned to parts of the inverted sequence sharing spurious sequence similarity with the uninverted sequences; or, there will be inference of an inserted sequence of unknown origin (in reality, the inverted nucleotides), which corresponds with a deletion in the homologous uninverted sequences. Each possibility will lead to inaccurate assessment of homology and may potentially have a considerable effect on phylogeny estimation.

Regions in the matrix demonstrating many independent variable-length insertion and deletion events will likely be associated with secondary structures, specifically with loop regions of stem-loops (Kelchner, 1996; Downie et al., 1996b; Kelchner & Clark, 1997). Identification of flanking sequences involved in possible stem formation could locate the boundaries for the region and aid in aligning the indels. Discerning probable SSM-susceptible sites can also be informative for the inference of parallel and reversed insertions or deletions.

Perhaps methods of gap or character weighting and alignment based on mechanisms of mutation can be incorporated into software designed for non-coding sequence alignment, particularly by including an evaluation of ΔG values for probable secondary structures. However, the diversity of rates and types of molecular evolution in non-coding regions may be profound. As with coding DNA, we are far from understanding all forces directing non-coding molecular evolution to a degree that we can, with any certainty, assign probabilities to individual mutations.

Considering that alignment of sequence data is fundamental to the entire phylogeny estimation process, authors should more fully describe the steps taken to align their sequence data in order to provide necessary information for the assessment of their proposed reconstructions of phylogenies.

Analysis of Non-Coding Sequence Data

The mechanisms of evolution described above have a number of significant implications for the phylogenetic analysis of non-coding sequence data. Among these are the following:

(1) Slipped-strand mispairing can be the result of persistent mutational triggers (especially when the trigger sequence is located in the stem of a stem-loop secondary structure). This can introduce homoplasy from parallelisms and reversals into any phylogenetic estimations that include gap-coded characters in the matrix. Multiple indel events in a localized region may obscure homology of length mutations. Non-independence of these mutations introduces the issue of relative weighting of nucleotide characters linked in a repeat unit, if each base is treated as a character in an analysis. Weight of the unit taken as a single character is also an issue if the unit is included in the analysis as a coded gap character.

(2) Secondary structure shows nonrandom mutation in the form of compensatory mutation and possible homogenization of sequence necessary for stem formation. Loop sequence is available for multiple mutations in the form of inversions, length mutations, and multiple-hit point substitutions, any of which may obscure evolutionary history.

(3) Inversions may show high levels of parallelism and reversal, and their phylogenetic utility may not be particularly robust. Undetected minute inversions may be buried within a data matrix and consequently treated as multiple base substitution synapomorphies instead of a single mutational event.

(4) Nucleotide substitutions may be under peculiar constraints not fully understood. There is evidence of a bias in non-coding regions involving transition/transversion substitution ratios due to the influence of neighboring bases. A particular base may experience substitution events multiple times in closely related lineages, reaching saturation long before the expected saturation level for the remaining sequence. A base-composition bias toward A/T content is clearly present in non-coding cpDNA.

Selective pressures exerted on non-coding regions may be largely a function of the physical structure of the sequence and possible functionality of introns and intergenic spacers. Reliance on methodology developed for coding sequence, which includes estimates of constraints on coding sequence evolution, transition/transversion ratios, and mutation probabilities, is inappropriate for the analysis of non-coding regions.

Phylogenetic estimations based on genetic distance measures of non-coding cpDNA sequences must be approached with care. Superficial application of models for maximum likelihood (ML; Felsenstein, 1981) or neighbor-joining (NJ; Saitou & Nei, 1987) could easily produce erroneous phylogenetic estimations if several key assumptions underlying the methodology are violated.

For example, most models consider a nucleotide site as the unit of evolution (Ritland & Eckenwalder, 1992), a consideration that is contradicted by the mode of non-coding sequence evolution. Simplistic models based on the commonly calculated Kimura estimates (Kimura, 1980) and Jukes-Cantor estimates (Jukes & Cantor, 1969) assume an equal 25% frequency for each nucleotide type throughout the sequence and generate base mutation probabilities from this assumption. Because non-coding cpDNA regions can demonstrate much higher A/T content, this assumption is clearly contradicted. Furthermore, transition/transversion ratios in non-coding regions can differ considerably from coding ones (see Hoot & Douglas, 1998), and may even vary between discrete non-coding regions of the chloroplast genome. Among-site mutation rate heterogeneity is highly probable, especially if regions of conservation and hot spots for mutation exist in the data. The presence of multiple gaps in an aligned matrix presents an additional hurdle for distance analysis, and indels themselves are difficult to incorporate as additional characters.

Countering such complications can be involved and computationally demanding. Modification of the initial Jukes-Cantor estimates to allow for varying base frequencies (e.g., Tajima & Nei, 1984) should be employed. Transition/transversion ratios can be estimated directly from the non-coding sequence matrix by pairwise sequence comparisons (e.g., Yang & Yoder, 1999), eliminating the circularity occasioned by measures derived from a topology. More refined distance models that incorporate these problems stand a better chance of reflecting the underlying manner of molecular evolution in non-coding sequence data. Such refined models may therefore estimate a more accurate phylogeny that better recovers the evolutionary history of the characters.

With ML, transition/transversion estimates are dependent on whether among-site rate variation has been incorporated in the model and can be sensitive to the accuracy of the topology used for their estimation (Sullivan et al., 1996). Among-site rate heterogeneity in the data is often assumed to fit either a negative binomial or gamma distribution function, and confirmation can be assessed statistically. Such rate heterogeneity is likely present in non-coding sequence data due to the effects of secondary structure on mutation likelihoods. Rates of variation at sites are usually expected to fit a gamma distribution model (Yang, 1996), and a parameter (α) can be determined to define the shape of that underlying function in an ML analysis (see Yang (1994) and Yang (1996) for thorough explanation). However, Sullivan et al. (1996) suggested α estimates are strongly affected by the topology used for their estimation. Therefore, to improve the ability of a model incorporating gamma distribution to recover the "correct" phylogeny, α must be calculated directly from the data matrix; this should be done by pairwise comparison, which can be a computationally intensive or even impossible procedure as the number of taxa increases in the matrix (Yang, 1996; Sullivan et al., 1996). Poor estimation of α can easily result in a misleading phylogenetic hypothesis (Yang, 1996; Sullivan et al., 1996).

Other problems associated with non-coding cpDNA sequence data may be very difficult to address. If at least some of the mutation in non-coding sequences occurs in linked units, then the non-independence of these nucleotide characters directly affects the subsequent analysis. At present, there is no reliable parameter estimate to incorporate such non-independent characters in a distance model. Most work on parameter estimates for models has been based on coding sequence observations, and thus may not reflect the unique aspects of molecular evolution in non-coding regions.

Determining probabilistic estimates for non-coding cpDNA mutations is, at this time, difficult; therefore, the accurate assessment of the underlying mode of evolution for maximum likelihood analysis may be impossible. As Yang et al. (1995) discussed in detail, the accuracy of ML in recovering an evolutionary history is strongly dependent on the evolutionary model applied. Thus, for non-coding cpDNA sequence data (as well as genic sequence data), deeper understanding of the manner of evolution in these regions is required before an accurate model for ML phylogenetic analysis can be applied.

The frequent alternative to distance measures and maximum likelihood is parsimony analysis. Heuristic parsimony searches can be considerably faster and less computationally intensive than a maximum likelihood analysis with the parameter adjustments described above; however, they are often much slower than a distance analysis. Parsimony analyses that contain no weighting schemes for transition/transversion bias and non-independent mutation of matrix characters may be as vulnerable to recovery of an inaccurate phylogeny as similarly simplistic distance models. It has been suggested that parsimony's potential in some cases to recover a correct topology decreases significantly when among-site rate heterogeneity exists in the data (Tateno et al., 1994; Kuhner & Felsenstein, 1994; Huelsenbeck, 1995). Such rate heterogeneity could arise from the structured sequence patterns described here in non-coding cpDNA. And though it has been proposed that the reliability of parsimony estimates increases with increasing number of taxa included in an analysis (e.g., Wakeley, 1993; Sullivan et al., 1995; Yang, 1996), it is unclear if this effect is independent of possible among-site rate variation.

Parsimony specifies no particular probabilistic evolutionary model, but like all phylogenetic estimation methods it is influenced by non-independence of characters. This problem can be alleviated to a degree if mutations such as inversions and inserted or deleted repeats are recognized as non-independent events and are either excluded from the analysis or coded separately as described below. Any non-independent evolution of neighboring nucleotides in a sequence would create an artificial weighting effect for these positions in a parsimony analysis that considers each nucleotide an independently evolving character.

Various weighting schemes have been proposed to counter this effect. Weighting has been applied, for example, to compensatory mutations associated with secondary structure in rDNA (e.g., Wheeler & Honeycutt, 1988; Dixon & Hillis, 1993; Baldwin et al., 1995; Soltis et al., 1997; Soltis & Soltis, 1998). Trial weighting schemes have also been applied to non-coding sequence data from the chloroplast (e.g., Downie et al., 1996a; Liden et al., 1997). However, Olmstead et al. (1998) reasoned that an erroneous weighting model increases the chance that the correct topology is excluded from the most parsimonious topologies recovered. In their opinion, a more general model such as equal weighting of characters may limit resolution, but would increase the chance that the "true" tree is recovered by the analysis. Development of defensible weighting schemes for non-coding sequence data would necessarily come from evidence provided by comparative analysis of non-coding regions throughout the chloroplast genome, and may be specific to individual data sets. The likelihood of misdiagnosing an appropriate weighting scheme for subsets of the data may still be high. Therefore, it is perhaps sensible for now to apply equal weighting to non-coding sequence characters until we have further evidence to support a particular weighting scheme.

Insertions and deletions have been shown to be of considerable phylogenetic value (e.g., Golenberg et al., 1993; Mes & Hart, 1994; Natali et al., 1995; Downie et al., 1996a; Kelchner & Clark, 1997; Oxelman et al., 1997; Sang et al., 1997; Liden et al., 1997; Downie et al., 1998; Bayer & Starr, 1998), and one should consider including gaps as coded (present/absent) characters appended to the sequence matrix (e.g., Hodges & Arnold, 1994; Kelchner & Clark, 1997; Sang et al., 1997; Downie et al., 1998; Hoot & Douglas, 1998; Bayer & Starr, 1998). Selection of gaps to be included in the analysis, however, is somewhat subjective in that optimally only those length mutations arguably homologous based on size, composition, and related mechanistic origin should be included.

The exclusion of gaps and removal of coded gap characters from a non-coding sequence matrix can be an interesting and informative approach to studying the degree of resolution provided by point substitution information alone (e.g., Kelchner, 1996; Kelchner & Clark, 1997). A similar analysis can be conducted by including coded gap characters only and excluding all other characters in the matrix. Coupled with mapping characters onto a topology produced from a complete matrix, these partitioned analyses may prove useful in locating and determining the degree of problematic homoplasy affecting resolution in competing topologies.

Minute inversions should be identified and removed from the analysis, to be added as present/absent characters at the end of the matrix (Kelchner & Wendel, 1996; Kelchner & Clark, 1997). This eliminates potential scoring of multiple non-homologous synapomorphies that are artifacts of an inversion mutation.

Of some concern is the tendency to treat nucleotide gap characters of taxa that do not share an insertion (i.e., have only spaces present at the insertion position in the matrix) as missing characters when conducting parsimony analysis. This results in inferred nucleotide homology for characters in the inserted sequences, which leads to cladistic assessment of their base substitutions. Such an approach should be applied only when evidence of the homology of inserted sequences is convincing. Chaotic regions or other areas where homology assessment is deemed impossible should be excluded from the data matrix before analysis (see Liden et al., 1997) to avoid this mistaken claim of nucleotide homology.

Bootstrap (Felsenstein, 1985) and jackknife (Farris et al., 1997) analyses, frequently misunderstood to be direct measures of phylogenetic accuracy, are only as sound as their underlying analysis procedure. As with coding sequences (see Trueman, 1993; Hillis & Bull, 1993; Bremer, 1994; Mishler, 1994; Brown, 1994), both support measures can be affected by the non-independent structure present in non-coding sequences. The structure invalidates a requirement of the statistic that each nucleotide be a discrete and independent character.

Bootstrap and jackknife analyses are a re-sampling of the data matrix in an effort to statistically measure how robustly the data in the matrix support a particular topology. The concept is sound, but the statistical integrity of both measures relies on the assumption that each nucleotide is an individual character, that each character evolves randomly and independently, and that the matrix represents a sample of a much larger population of characters evolving in identical fashion (Felsenstein, 1985). Due to the non-independent structure existing in non-coding regions, and the probably unique series of evolutionary constraints acting not only on individual non-coding regions but also on partitions of a region, each of these assumptions may be violated. Sampling from within such a data set equates to sampling a nonrandom and non-independent subset of a non-existing larger population. A large number of bootstrap replicates should, in theory, cover all possible error due to reduced character sampling in each replicate, but the strength of the bootstrap test is weakened if the characters are not accurately defined. If a character in some cases is not an individual nucleotide but a suite of nucleotides, the conditions that would make bootstrapping and jackknifing accurate as measures of data support for a topology are not satisfied. An analysis would produce an unequal weighting effect on subsets of the data in each resampling due to the frequent localized violation of character definition.

A non-resampling technique that allows assessment of data support for individual clades is the Bremer Support measure (BS, or "decay" analysis; Bremer, 1988, 1994; Donoghue et al., 1992; for application to large data sets, see Baum et al., 1994; Morgan, 1997). The measure is a function only of the recoverability of clades in topologies progressively one step longer. Bremer support has the possibility of sidestepping the effects of character definition issues discussed above for bootstrapping if the model underlying the phylogeny estimation considers the variable nature of character definition in a nucleotide set.

Oxelman et al. (1999) demonstrated that bootstrapping and BS evaluate different parameters of the data matrix, and are thus not directly comparable measures (though BS values, when high, may be imperfectly correlated with bootstrap and jackknife values). BS values cannot be viewed as probabilistic estimates themselves (Oxelman et al., 1999), and an inability to adapt the measures to a standard scale that is universally applicable renders the technique of dubious worth to some systematists. However, the innovation by Oxelman et al. (1999) that includes minimal branch length values with each BS value does, in a non-standard way, improve the comparative information capacity of the measure. This procedure may be more meaningful and informative than bootstrap and jackknife values for non-coding cpDNA data.

Conclusions

In summary, great care should be given to the alignment and assessment of non-coding sequence data. There is considerable evidence now that non-coding regions are highly structured, non-randomly evolving DNA; thus, alignment by current randomized algorithmic software is rarely adequate. An understanding of the proposed mechanisms of mutation acting on non-coding sequences is critical for the positioning of gaps and the better assessment of homology of indels and point substitutions. Probable secondary structure should be routinely identified and used as an important source of information to aid in aligning chaotic or labile regions of the data matrix. Prior to phylogenetic analysis, all matrices should be carefully reviewed for obscured mutational events, such as minute inversions or misaligned repeat units.

Important for understanding molecular evolution in non-coding DNA is the concept of the mutational trigger (Kelchner, 1996; Kelchner & Clark, 1997), a specific sequence pattern that creates the foundation for a mutational event. Such triggers often remain intact after generating a mutation, and their presence can easily occasion a repeated, paralleled, or reversed mutation event. Triggers may likely be responsible for much of the homoplasy of gap characters inferred in studies at any taxonomic level; those applying non-coding sequence data to molecular systematics should be aware of their occurrence and effect.

Information of the kind presented here can increase the predictive value of mutational events in non-coding DNA. For example, Kelchner and Wendel (1996) suggested that minute inversions associated with hairpin secondary structures described in non-coding cpDNA could occur in similar situations in other genomes. Dumolin-Lapègue et al. (1998) recently reported just such an event in the mitochondria of oak populations of southern France. Hence, recommendations proposed in this paper for the phylogenetic analysis of non-coding cpDNA sequences may likely apply to data from non-coding regions of nuclear, and particularly mitochondrial, genomes.

Choosing an appropriate non-coding region for a particular taxonomic level is essential for maximizing its utility as a phylogenetic tool, but there is no infallible method for determining what that "proper" degree of mutation is for a particular study. A region's utility may vary between plant groups that are assumed to occupy the same evolutionary level, and data from multiple non-coding regions, when applied to one taxonomic group, can vary remarkably in phylogenetic utility (see Small et al., 1998). In light of the mutational mechanisms outlined in this article, at least one concern seems justified: if the taxonomic level is too high, one would expect saturation of multiple hit sites and concealment of multiple hit indels in any non-coding region, decreasing its utility as a phylogenetic tool.

The perceived intricacies of molecular evolution and their bearing on phylogenetic analysis, both in non-coding and coding regions (for genes have well-known mechanistic biases as well—the codon position being just one example) can be discouraging. However, the phenomena outlined in this article have solutions in most cases, and attention to alignment and analysis should enhance the phylogenetic utility and accuracy of non-coding cpDNA data. It should be noted that in almost all systematic studies based on non-coding cpDNA sequences, the authors profess to have found sufficient phylogenetic information in their data to warrant its use in lower-level phylogenetic analyses.

Clearly there is a need to develop an understanding of molecular evolution in non-coding cpDNA regions similar to that which exists for chloroplast genic DNA. Continued research into non-coding sequence evolution may eventually produce a more balanced process for the alignment and phylogenetic analysis of non-coding sequence data. Future software may be able to measure and assess probabilities associated with particular mutational mechanisms and incorporate this information into the alignment process. This would be an immense aid to those systematists who wish to apply non-coding molecular tools to the field of plant systematics.

Literature Cited

Aldrich, J., B. W. Cherney, E. Merlin & L. Christopherson. 1988. The role of insertion/deletions in the evolution of the intergenic region between *pbs*A and *trn*H in the chloroplast genome. Curr. Genet. 14: 137–146.

Asmussen, C. B. & A. Liston. 1998. Chloroplast DNA characters, phylogeny, and classification of *Lathyrus* (Fabaceae). Amer. J. Bot. 85: 387–401.

Baldwin, B. G., M. J. Sanderson, J. M. Porter, M. F. Wojciechowski, C. S. Campbell & M. J. Donoghue. 1995. The ITS region of nuclear ribosomal DNA: A valuable source of evidence on angiosperm phylogeny. Ann. Missouri Bot. Gard. 82: 247–277.

Baum, D. A., R. L. Small & J. F. Wendel. 1998. Biogeography and floral evolution of baobabs (*Adansonia*, Bombacaceae) as inferred from multiple data sets. Syst. Biol. 47: 181–207.

———, K. J. Sytsma & P. C. Hoch. 1994. A phylogenetic analysis of *Epilobium* (Onagraceae) based on nuclear ribosomal DNA sequences. Syst. Bot. 19: 363–388.

Bayer & J. R. Starr. 1998. Tribal phylogeny of the Asteraceae based on two non-coding chloroplast sequences, the *trn*L intron and *trn*L/*trn*F intergenic spacer. Ann. Missouri Bot. Gard. 85: 242–256.

———, C. F. Puttock & S. A. Kelchner. 2000. Phylogeny of South African Gnaphalieae (Asteraceae) based on two non-coding chloroplast sequences. Amer. J. Bot. 87: 259–272.

Blasko, K., S. A. Kaplan, K. G. Higgins, R. Wolfson & B. B. Sears. 1988. Variation in copy number of a 24-base pair tandem repeat in the chloroplast DNA of *Oenothera hookeri* strain Johansen. Curr. Genet. 14: 287–292.

Böhle, U.-R., H. Hilger, R. Cerff & W. F. Martin. 1994. Non-coding chloroplast DNA for plant molecular systematics at the infrageneric level. Pp. 391–403 *in* B. Schierwater, B. Streit, G. P. Wagner & R. DeSalle (editors), Molecular Ecology and Evolution: Approaches and Applications. Birkhäuser Verlag, Basel.

———, ——— & W. F. Martin. 1997. Island colonization and evolution of the insular woody habit in *Echium* L. (Boraginaceae). Proc. Natl. Acad. Sci. U.S.A. 93: 11740–11745.

Bonnard, G., F. Michel, J. H. Weil & A. Steinmetz. 1984. Nucleotide sequence of the split *t*RNA(Leu/UAA) gene from *Vicia faba* chloroplasts: Evidence for structural homologies of the chloroplast *t*RNA(Leu) intron with the intron from the autosplicable *Tetrahymena* ribosomal RNA precursor. Molec. Gen. Genet. 194: 330–336.

Bremer, K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. Evolution 42: 795–803.

———. 1994. Branch support and tree stability. Cladistics 10: 295–304.

Brown, J. K. M. 1994. Bootstrap hypothesis tests for evolutionary trees and other dendrograms. Proc. Natl. Acad. Sci. U.S.A. 91: 12293–12297.

Buroker, N. E., J. R. Brown, T. A. Gilbert, P. J. O'Hara, A. T. Beckenback, W. K. Thomas & M. J. Smith. 1990. Length heteroplasmy of sturgeon mitochondrial DNA: An illegitimate elongation model. Genetics 124: 157–163.

Cech, T. R. 1990. Self-splicing of Group I introns. Annual Rev. Biochem. 59: 543–568.

Clegg, M. T. & G. Zurawski. 1992. Chloroplast DNA and the study of plant phylogeny: Present status and future prospects. Pp. 1–13 *in* P. S. Soltis, D. E. Soltis & J. J. Doyle (editors), Molecular Systematics of Plants. Chapman and Hall, New York.

———, K. Ritland & G. Zurawski. 1986. Processes of chloroplast DNA evolution. Pp. 275–294 *in* S. Karlin & E. Nevo (editors), Evolutionary Processes and Theory. Academic Press, New York.

———, B. S. Gaut, G. H. Learn, Jr. & B. R. Morton. 1994. Rates and patterns of chloroplast DNA evolution. Proc. Natl. Acad. Sci. U.S.A. 91: 6795–6801.

Cros, J., M. C. Combes, P. Trouslot, F. Anthony, S. Hamon, A. Charrier & P. Lashermes. 1998. Phylogenetic analysis of chloroplast DNA variation in *Coffea* L. Molec. Phylogenet. Evol. 9: 109–117.

Cummings, M. P., L. M. King & E. A. Kellogg. 1994. Slipped-strand mispairing in a plastid gene: *rpo*C2 in grasses (Poaceae). Molec. Biol. Evol. 11: 1–8.

Curtis, S. E. & M. T. Clegg. 1984. Molecular evolution of chloroplast DNA sequences. Molec. Biol. Evol. 1: 291–301.

Curtiss, W. C. & J. N. Vournakis. 1984. Quantification of base substitutions in eukaryotic 5S rRNA: Selection for the maintenance of RNA secondary structure. J. Molec. Evol. 20: 351–361.

Dixon, M. T. & D. M. Hillis. 1993. Ribosomal RNA secondary structure: Compensatory mutations and implications for phylogenetic analysis. Molec. Biol. Evol. 10: 256–267.

Donoghue, M. J., R. G. Olmstead, J. F. Smith & J. D. Palmer. 1992. Phylogenetic relationships of Dipsacales based on *rbc*L sequences. Ann. Missouri Bot. Gard. 79: 333–345.

Downie, S. R., D. S. Katz-Downie & K.-J. Cho. 1996a. Phylogenetic analysis of Apiaceae subfamily Apioideae using nucleotide sequences from the chloroplast *rpoC1* intron. Molec. Phylogenet. Evol. 6: 1–18.

———, E. Llanas & D. S. Katz-Downie. 1996b. Multiple independent losses of the *rpo*C1 intron in angiosperm chloroplast DNA's. Syst. Bot. 21: 135–151.

———, S. Ramanath, D. S. Katz-Downie & E. Llanas. 1998. Molecular systematics of Apiaceae subfamily Apioideae: Phylogenetic analyses of nuclear ribosomal DNA internal transcribed spacer and plastid *rpo*C1 intron sequences. Amer. J. Bot. 85: 563–591.

Dujon, B. 1989. Group I introns as mobile genetic elements: Facts and mechanistic speculations—A review. Gene 82: 91–114.

Dumolin-Lapègue, S., M.-H. Pemonge & R. J. Petit. 1998. Association between chloroplast and mitochondrial lineages in oaks. Molec. Biol. Evol. 15: 1321–1331.

Farris, J. S., V. A. Albert, M. Kallersjö, D. Lipscomb & A. G. Kluge. 1997. Parsimony jackknifing outperforms neighbor-joining. Cladistics 12: 99–124.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Molec. Evol. 17: 368–376.

———. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39: 783–791.

Ferris, C., R. P. Oliver, A. J. Davy & G. M. Hewitt. 1995. Using chloroplast DNA to trace postglacial migration routes of oaks into Britain. Molec. Ecol. 4: 731–738.

Gielly, L. & P. Taberlet. 1994. The use of chloroplast DNA to resolve plant phylogenies: Non-coding versus *rbc*L sequences. Molec. Biol. Evol. 11: 769–777.

Golenberg, E. M., M. T. Clegg, M. L. Durbin, J. Doebley & D. P. Ma. 1993. Evolution of a non-coding region of the chloroplast genome. Molec. Phylogenet. Evol. 2: 52–64.

Ham, R. C. H. J. van, H. t'Hart, T. H. M. Mes & J. M. Sandbrink. 1994. Molecular evolution of non-coding regions of the chloroplast genome in the Crassulaceae and related species. Curr. Genet. 25: 558–566.

Hancock, J. M. 1995. The contribution of slippage-like processes to genome evolution. J. Molec. Evol. 41: 1038–1047.

Hibbett, D. S. 1996. Phylogenetic evidence for horizontal transmission of Group I introns in the nuclear ribosomal DNA of mushroom-forming fungi. Molec. Biol. Evol. 13: 903–917.

Higgins, D. G., A. J. Bleasby & R. Fuchs. 1992. CLUSTAL V: Improved software for multiple sequence alignment. Computer Applic. Biosci. 8: 189–191.

Hillis, D. M. & J. J. Bull. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42: 182–192.

Hodges, S. A. & M. L. Arnold. 1994. Columbines: A geographically widespread species flock. Proc. Natl. Acad. Sci. U.S.A. 91: 5129–5132.

Hoot, S. B. & A. W. Douglas. 1998. Phylogeny of the Proteaceae based on *atp*B and *atp*B-*rbc*L intergenic spacer region sequences. Austral. Syst. Bot. 11: 301–320.

——— & J. D. Palmer. 1994. Structural rearrangements, including parallel inversions, within the chloroplast genome of *Anemone* and related genera. J. Molec. Evol. 38: 274–281.

Howe, C. J. 1985. The endpoints of an inversion in wheat chloroplast DNA are associated with short repeated sequences containing homology to *att*-lambda. Curr. Genet. 10: 139–145.

Huelsenbeck, J. P. 1995. The performance of phylogenetic methods in simulation. Syst. Biol. 44: 17–48.

Jaeger, J. A., D. H. Turner & M. Zuker. 1989. Improved predictions of secondary structures for RNA. Proc. Natl. Acad. Sci. U.S.A. 86: 7706–7710.

Johnson, L. A. & D. E. Soltis. 1994. *mat*K DNA sequences and phylogenetic reconstruction in Saxifragaceae *s. str.* Syst. Bot. 19: 143–156.

Jordan, W. C., M. W. Courtney & J. E. Neigel. 1996. Low levels of intraspecific genetic variation at a rapidly evolving chloroplast DNA locus in North American duckweeds (Lemnaceae). Amer. J. Bot. 83: 430–439.

Jukes, T. H. & C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. Munro (editor), Mammalian Protein Metabolism. Academic Press, New York.

Kajita, T., K. Kamiya, K. Nakamura, H. Tachida, R. Wickneswari, Y. Tsumura, H. Yoshimaru & T. Yamazaki. 1998. Molecular phylogeny of Dipterocarpaceae in Southeast Asia based on nucleotide sequences of *mat*K, *trn*L intron, and *trn*L-*trn*F intergenic spacer region in chloroplast DNA. Molec. Phylogenet. Evol. 10: 202–209.

Kanno, A. & A. Hirai. 1992. Comparative studies of the structure of chloroplast DNA from four species of *Oryza*: Cloning and physical maps. Theor. Appl. Genet. 83: 791–798.

———, N. Watanabe, I. Nakamura & A. Hirai. 1993. Variations in chloroplast DNA from rice (*Oryza sativa*): Differences between deletions mediated by short direct-repeat sequences within a single species. Theor. Appl. Genet. 86: 579–584.

Kelchner, S. A. 1996. Molecular evolution and phylogenetic utility of the chloroplast *rpl*16 intron in *Chusquea* and the Bambusoideae (Poaceae). M.Sc. Thesis, Department of Botany, Iowa State University, Ames, Iowa.

——— & L. G. Clark. 1997. Molecular evolution and phylogenetic utility of the chloroplast *rpl*16 intron in *Chusquea* and the Bambusoideae (Poaceae). Molec. Phylogenet. Evol. 8: 385–397.

——— & J. F. Wendel. 1996. Hairpins create minute inversions in non-coding regions of chloroplast DNA. Curr. Genet. 30: 259–262.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Molec. Evol. 16: 111–120.

Kohchi, T., Y. Ogura, K. Umesono, Y. Yamada, T. Komano, H. Ozeki & K. Ohyama. 1988. Ordered processing and splicing in a polycistronic transcript in liverwort chloroplasts. Curr. Genet. 14: 147–154.

Kuhner, M. K. & J. Felsenstein. 1994. A simulation of phylogeny algorithms under equal and unequal evolutionary rates. Molec. Biol. Evol. 11: 459–468.

Learn, G. H., Jr., J. S. Shore, G. R. Furnier, G. Zurawski & M. T. Clegg. 1992. Constraints on the evolution of plastid introns: The group II intron in the gene encoding *t*RNA-Val(UAC). Molec. Biol. Evol. 9: 856–871.

Levinson, G. & G. A. Gutman. 1987. Slipped-strand mispairing: A major mechanism for DNA sequence evolution. Molec. Biol. Evol. 4: 203–221.

Liden, M., T. Fukuhara, J. Rylander & B. Oxelman. 1997. Phylogeny and classification of Fumariaceae, with emphasis on *Dicentra* s.l., based on the plastid gene *rps*16 intron. Pl. Syst. Evol. 206: 411–420.

Mes, T. H. M. & H. t'Hart. 1994. *Sedum surculosum* and *S. jaccardianum* (Crassulaceae) share a unique 70bp deletion in the chloroplast DNA *trn*L(UAA)-*trn*F(GAA) intergenic spacer. Pl. Syst. Evol. 193: 213–221.

Michel, F. & B. Dujon. 1983. Conservation of RNA secondary structures in two intron families including mitochondrial-, chloroplast-, and nuclear-encoded members. EMBO J. 2: 33–38.

——— & E. Westhof. 1990. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. J. Molec. Biol. 216: 585–610.

———, K. Umesono & H. Ozeki. 1989. Comparative and functional anatomy of group II catalytic introns—A review. Gene 82: 5–30.

Milligan, B. G., J. N. Hampton & J. D. Palmer. 1989. Dispersed repeats and structural reorganization in subclover chloroplast DNA. Molec. Biol. Evol. 6: 355–368.

Mishler, B. D. 1994. Cladistic analysis of molecular and morphological data. Amer. J. Phys. Anthropol. 94: 143–156.

Morgan, D. R. 1997. Decay analysis of large sets of phylogenetic data. Taxon 46: 509–517.

Morton, B. R. 1995a. Neighboring base composition and transversion/transition bias in a comparison of rice and maize chloroplast non-coding regions. Proc. Natl. Acad. Sci. U.S.A. 92: 9717–9721.

———. 1995b. Neighboring base composition is strongly correlated with base substitution bias in a region of the chloroplast genome. J. Molec. Evol. 41: 597–603.

——— & M. T. Clegg. 1993. A chloroplast DNA mutational hotspot and gene conversion in a non-coding region near *rbc*L in the grass family (Poaceae). Curr. Genet. 24: 357–365.

———, V. M. Oberholzer & M. T. Clegg. 1997. The influence of specific neighboring bases on substitution bias in noncoding regions of the plant chloroplast genome. J. Molec. Evol. 45: 227–231.

Natali, A., J.-F. Manen & F. Ehrendorfer. 1995. Phylogeny of the Rubiaceae–Rubioideae, in particular the tribe Rubieae: Evidence from a non-coding chloroplast DNA sequence. Ann. Missouri Bot. Gard. 82: 428–439.

Ogihara, Y., T. Terachi & T. Sasakuma. 1988. Intramolecular recombination of chloroplast genome mediated by short direct-repeat sequences in wheat species. Proc. Natl. Acad. Sci. U.S.A. 85: 8573–8577.

Olmstead, R. G. & J. D. Palmer. 1994. Chloroplast DNA systematics: A review of methods and data analysis. Amer. J. Bot. 81: 1205–1224.

———, P. A. Reeves & A. C. Yen. 1998. Patterns of sequence evolution and implications for parsimony analysis of chloroplast DNA. Pp. 164–187 *in* D. S. Soltis, P. S. Soltis & J. J. Doyle (editors), Molecular Systematics of Plants II: DNA Sequencing. Chapman and Hall, New York.

Oxelman, B., M. Backlund & B. Bremer. 1999. Relationships of the Buddlejaceae s.l. investigated using parsimony jackknife and branch support analysis of chloroplast *ndh*F and *rbc*L sequence data. Syst. Bot. 24: 164–182.

———, M. Liden & D. Berglund. 1997. Chloroplast *rps*16 intron phylogeny of the tribe Sileneae (Caryophyllaceae). Pl. Syst. Evol. 206: 393–410.

Palmer, J. D. 1985. Comparative organization of chloroplast genomes. Annual Rev. Genet. 19: 325–354.

———. 1987. Chloroplast DNA evolution and biosystematic uses of chloroplast DNA variation. Amer. Naturalist 130 (suppl.): S6–S29.

———. 1991. Plastid chromosomes: Structure and evolution. Pp. 5–53 *in* L. Bogorad & I. K. Vasil (editors), Cell Culture and Somatic Cell Genetics of Plants, Vol. 7A: The Molecular Biology of Plastids. Academic Press, Orlando, Florida.

———, R. A. Jorgensen & W. F. Thompson. 1985. Chloroplast DNA variation and evolution in *Pisum*: Patterns of change and phylogenetic analysis. Genetics 109: 195–213.

———, J. M. Nugent & L. A. Herbon. 1987. Unusual structure of geranium chloroplast DNA: A triple-sized inverted repeat, extensive gene duplications, multiple inversions, and two repeat families. Proc. Natl. Acad. Sci. U.S.A. 84: 769–773.

Peterson, G. & O. Seberg. 1997. Phylogenetic analysis of the Triticeae (Poaceae) based on *rpo*A sequence data. Molec. Phylogenet. Evol. 7: 217–230.

Rigaa, A., M. Monnerot & D. Sello. 1995. Molecular cloning and complete nucleotide sequence of the repeated unit and flanking gene of the scallop *Pecten maximus* mitochondrial DNA: Putative replication origin features. J. Molec. Evol. 41: 189–195.

Ritland, K. & J. E. Eckenwalder. 1992. Polymorphism, hybridization, and variable evolutionary rate in molecular phylogenies. Pp. 404–428 *in* P. S. Soltis, D. E. Soltis & J. J. Doyle (editors), Molecular Systematics of Plants. Chapman and Hall, New York.

Rychlik, W. & R. E. Rhoads. 1989. A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA. Nucleic Acids Res. 17: 8543–8551.

Saitou, N. & M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molec. Biol. Evol. 4: 406–425.

Samuel, R., W. Pinsker & M. Kiehn. 1997. Phylogeny of some species of *Cyrtandra* (Gesneriaceae) inferred from the *atp*B/*rbc*L cpDNA intergene region. Bot. Acta. 110: 503–510.

Sang, T., D. J. Crawford & T. F. Stuessy. 1997. Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). Amer. J. Bot. 84: 1120–1136.

Savolainen, V., R. Spichiger & J.-F. Manen. 1997. Polyphyletism of Celastrales deduced from a chloroplast noncoding DNA region. Molec. Phylogenet. Evol. 7: 145–157.

Schnabel, A. & J. F. Wendel. 1998. Cladistic biogeography of *Gleditsia* (Leguminosae) based on *ndh*F and *rpl*16 chloroplast gene sequences. Amer. J. Bot. 85: 1753–1765.

Setoguchi, H., M. Ono, Y. Doi, H. Koyama & M. Tsuda. 1997. Molecular phylogeny of *Nothofagus* (Nothofagaceae) based on the *atp*B-*rbc*L intergenic spacer of the chloroplast DNA. J. Pl. Res. 110: 469–484.

Shimada, H. & M. Sugiura. 1991. Fine structural features of the chloroplast genome: Comparison of the sequenced chloroplast genomes. Nucleic Acids Res. 19: 983–995.

Small, R. L., J. A. Ryburn, R. C. Cronn, T. Seelanan & J. F. Wendel. 1998. The tortoise and the hare: Choosing between noncoding plastome and nuclear ADH sequences for phylogeny reconstruction in a recently diverged plant group. Amer. J. Bot. 85: 1301–1315.

Soltis, P. S. & D. E. Soltis. 1998. Molecular evolution of 18S rDNA in angiosperms: Implications for character weighting in phylogenetic analysis. Pp. 188–210 *in* D. S. Soltis, P. S. Soltis & J. J. Doyle (editors), Molecular Systematics of Plants II: DNA Sequencing. Chapman and Hall, New York.

———, L. A. Johnson & C. Looney. 1996. Discordance between ITS and chloroplast topologies in the *Boykinia* group (Saxifragaceae). Syst. Bot. 21: 169–185.

———, P. S. Soltis, D. L. Nickrent, L. A. Johnson, W. J.

Hahn, S. B. Hoot, J. A. Sweere, R. K. Kuzoff, K. A. Kron, M. W. Chase, S. M. Swensen, E. A. Zimmer, S.-M. Chaw, L. J. Gillespie, W. J. Kress & K. J. Sytsma. 1997. Angiosperm phylogeny inferred from 18S ribosomal DNA sequences. Ann. Missouri Bot. Gard. 84: 1–49.

Stein, J. vom & W. Hatchel. 1988. Deletions/insertions, short inverted repeats, sequences resembling *att*-lambda, and frame-shifted-mutated open reading frames are involved in chloroplast DNA differences in the genus *Oenothera* subsection Munzia. Molec. Gen. Genet. 213: 513–518.

Streisinger, G. & J. Owen. 1985. Mechanisms of spontaneous and induced frameshift mutation in bacteriophage T4. Genetics 109: 633–659.

Sullivan, J., K. E. Holsinger & C. Simon. 1995. Among-site rate variation and phylogenetic analysis of 12S rRNA in Sigmodontine rodents. Molec. Biol. Evol. 12: 988–1001.

———, ——— & ———. 1996. The effect of topology on estimates of among-site rate variation. J. Molec. Evol. 42: 308–312.

Tajima, F. & M. Nei. 1984. Estimation of evolutionary distance between nucleotide sequences. Molec. Biol. Evol. 1: 269–285.

Tateno, Y., N. Takezaki & M. Nei. 1994. Relative efficiencies of the maximum likelihood, neighbor joining, and maximum parsimony methods when substitution rate varies with site. Molec. Biol. Evol. 11: 261–277.

Thompson, J. D., D. G. Higgins & T. J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673–4680.

Trueman, J. W. H. 1993. Randomization confounded: A response to Carpenter. Cladistics 9: 101–109.

Wakeley, J. 1993. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J. Molec. Evol. 37: 613–623.

Wheeler, W. C. & R. L. Honeycutt. 1988. Paired sequence difference in ribosomal RNAs: Evolutionary and phylogenetic implications. Molec. Biol. Evol. 5: 90–96.

———, J. C. Gatesy & R. Desalle. 1995. Elision: A method for accommodating multiple molecular sequence alignments with alignment-ambiguous sites. Molec. Phylogenet. Evol. 4: 1–9.

Wolfe, K. H., W.-H. Li & P. M. Sharp. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. U.S.A. 84: 9054–9058.

Wolfson, R., K. G. Higgins & B. B. Sears. 1991. Evidence for replication slippage in the evolution of *Oenothera* chloroplast DNA. Molec. Biol. Evol. 8: 709–720.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Molec. Evol. 39: 306–314.

———. 1996. Among-site rate variation and its impact on phylogenetic analyses. Tree 11(9): 367–372.

——— & A. D. Yoder. 1999. Estimation of the transition/transversion rate bias and species sampling. J. Molec. Evol. 48: 274–283.

———, N. Goldman & A. Friday. 1995. Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. Syst. Biol. 44: 384–399.

Zuker, M. 1989. On finding all suboptimal foldings of an RNA molecule. Science 244: 48–52.

Zurawski, G. & M. T. Clegg. 1987. Evolution of higher-plant chloroplast DNA-encoded genes: Implications for structure-function and phylogenetic studies. Annual Rev. Pl. Physiol. 38: 398–418.